Over structurering van beoordelingsmethoden voor open vragen

SAMENVATTING

In de afgelopen dertig jaar is een groot scala aan instrumenten voor medisch probleemoplossen ontwikkeld. Deze instrumenten worden gekenmerkt door het gebruik van patiëntproblemen (casus). De kwaliteit van deze instrumenten wordt voor een belangrijk deel beperkt door het feit dat medisch probleemoplossen sterk inhoudsafhankelijk blijkt te zijn. In de recent ontwikkelde instrumenten wordt getracht aan dit inhoudsspecificiteitsprobleem tegemoet te komen door een groot aantal casus in een toets op te nemen. Dit kan worden bereikt door de vragen te beperken tot de essentiële aspecten (Bordage & Page, 1987; De Graaff, 1989; Page et al., 1990).
Bij deze instrumenten worden geregeld varianten van de gestructureerde open vraag gehanteerd (Feletti & Smith, 1986). De betrouwbaarheid van de beoordelingen vormt bij open vragen echter een probleem. Empirisch onderzoek heeft aangetoond dat de geringe overeenstemming tussen beoordelaars een belangrijke bron van onbetrouwbaarheid is. De mate waarin dergelijke beoordelaarseffecten optreden lijkt voor een belangrijk deel te worden beïnvloed door de kwaliteit van de gehanteerde beoordelingsmethode. Onderzoeken in het taalonderwijs laten bijvoorbeeld zien dat zowel een te sterke als te zwakke structurering van het beoordelingsproces onbetrouwbaarheid in de hand werkt. De precieze relatie tussen de mate van structurering en het optreden van beoordelaarseffecten is echter onduidelijk, evenals de rol die de inhoudsdeskundigheid van de beoordelaar hierin speelt.
In deze dissertatie wordt een experimenteel onderzoek beschreven waarin beide aspecten werden onderzocht. Daarnaast werd in het onderzoek aandacht besteed aan de invloed van (positieve) weging op validiteit en betrouwbaarheid van gegeven beoordelingen. Tevens werd de praktische bruikbaarheid van de beoordelingsmethoden geïnventariseerd.

Voor het onderzoek werden acht casus ontwikkeld, die medisch inhoudelijk waren gericht op de huisartspraktijk. Bij elke casus werden één tot drie korte antwoordvragen geformuleerd, die betrekking hadden op de essentie van het patiëntprobleem. De casus en bijbehorende vragen werden aan geneeskunde studenten uit verschillende jaargroepen ter beantwoording voorgelegd. Voor het beoordelen van de antwoorden werden per vraag vier beoordelingsmethoden ontwikkeld. Deze methoden varieerden van geen tot vergaande structurering wat betreft inhoudelijke richtlijnen. In volgorde van toenemende structurering waren dit: (1) het vrije oordeel, (2) de korte antwoordsleutel, (3) de ingedikte criterialijst en (4) de uitgebreide criterialijst. Bij de drie gestructureerde beoordelingsmethoden werden wegingsfactoren aangebracht om de invloed van positieve weging op betrouwbaarheid en validiteit te kunnen bestuderen.
Om inzicht te krijgen in de interactie tussen de mate van structurering en de inhoudsdeskundigheid van de beoordelaars, werden de antwoorden aan drie verschillende groepen beoordelaars voorgelegd. Deze beoordelaarsgroepen bestonden respectievelijk uit huisartsen, vierdejaars geneeskunde en vierdejaars economie studenten. De waardering van de beoordelaars voor de verschillende beoordelingsmethoden werd door middel van een vragenlijst onderzocht.

De theoretische validiteit van de casus werd onderzocht door de gemiddeld behaalde score

door de drie jaargroepen met elkaar te vergelijken. Indien de gemiddelde score toenam naarmate de studenten meer medisch onderwijs hadden gevolgd, werd dit opgevat als een indicatie voor theoretische validiteit.
De validiteit van de beoordelingsmethoden werd in eerste instantie onderzocht door middel van een correlationeel onderzoek, waarbij de correlatie tussen de verschillende beoordelingsmethoden werd berekend. Tevens werd de validiteit van de beoordelingsmethoden bestudeerd aan de hand van een discriminant-analyse. Hierbij werden de studenten op basis van de behaalde score op de casustoets in een jaargroep geplaatst en werd deze voorspelde indeling vergeleken met de daadwerkelijke groepsindeling van de studenten. Het percentage juiste classificaties werd als maat voor de validiteit genomen; hoe hoger het percentage des te meer valide is de beoordelingsmethode.

Voor de betrouwbaarheidsanalyses werd gebruik gemaakt van de generaliseerbaarheidstheorie. Om de relatieve bijdrage van de beoordelaars, de casus, de studenten en de interactie tussen deze factoren op de totale variantie te schatten werd per beoordelingsmethode een generaliseerbaarheidsstudie (G-studie) uitgevoerd. In de G-studie werd per beoordelingsmethode en voor elke groep beoordelaars een volledig gekruist p x i x j-design gehanteerd. Op basis van de G-studie werd een aantal D-studies verricht. In de eerste D-studie werd de interbeoordelaarsbetrouwbaarheid geschat. In de resterende D-studies werd de invloed van het aantal casus en beoordelaars op de totale betrouwbaarheid geschat. Deze D-studies werden uitgevoerd vanuit een norm en een criterium georiënteerd perspectief, gebruikmakend van zowel een volledig gekruist p x i x j-design als een genest p x j:i-design. Ten slotte werd een D-studie uitgevoerd waarin de consequenties van de gehanteerde beoordelingsmethode op de betrouwbaarheid van de genomen zak/slaag-beslissing werd bestudeerd. Hiertoe werd bij elke beoordelingsmethode de betrouwbaarheid berekend bij een cesuur van 30%, 40%, 50% en 60%.

De invloed van positieve weging werd bestudeerd door de gewogen scores te correleren met de ongewogen scores per beoordelingsmethode. Daarnaast werden de correlaties tussen de verschillende beoordelingsmethoden op basis van de ongewogen scores berekend. Deze resultaten werden vergeleken met de correlaties tussen de gewogen scores.

Voorafgaande aan het hoofdonderzoek werd een pilot-onderzoek uitgevoerd. Hierin werden de beoordelingsmethoden ontwikkeld en de wegingsfactoren empirisch bepaald. Tevens werd de validiteit van het onderzoeksmateriaal en de praktische uitvoerbaarheid van het onderzoeksdesign bestudeerd.
Nadat minimale veranderingen in het onderzoeksmateriaal waren aangebracht, werd op basis van het pilot-onderzoek besloten het hoofdonderzoek uit te voeren.

Overeenkomstig het pilot-onderzoek werd in het hoofdonderzoek gevonden dat de gemiddelde score op de casustoets toenam naarmate de studenten meer medisch onderwijs hadden gevolgd. Daarnaast nam bij elke jaargroep de gemiddelde score af naarmate de beoordelingsmethode meer was gestructureerd. Deze tendens was bij alle beoordelingsmethoden waarneembaar onafhankelijk van de medische expertise van de beoordelaars.
Wat betreft de correlaties tussen de verschillende beoordelingsmethoden werd het volgende geconstateerd: bij elke beoordelaarsgroep was de correlatie tussen het vrije oordeel en de gestructureerde beoordelingsmethoden kleiner dan de correlatie tussen de

gestructureerde beoordelingsmethoden onderling. Bij de economie studenten werden eveneens lagere correlaties gevonden tussen de korte antwoordsleutel en de twee meer gestructureerde beoordelingsmethoden. De correlatie tussen de ingedikte en de uitgebreide criterialijst varieerde bij de drie groepen beoordelaars tussen .75 (huisartsen) en .87 (geneeskunde studenten). Ook na de correctie voor attenuatie, uitgevoerd met de interbeoordelaarsbetrouwbaarheid, bleef hetzelfde beeld gehandhaafd.
De validiteit van de beoordelingsmethoden werd vervolgens bestudeerd aan de hand van een discriminant-analyse. Uit deze analyse kwamen geen duidelijke verschillen tussen de beoordelingsmethoden naar voren. Elke groep beoordelaars bereikte met elke beoordelingsmethode een percentage van juiste classificaties tussen 76% en 80%.

Per beoordelingsmethode werd de correlatie berekend tussen de drie groepen beoordelaars. Bij het vrije oordeel werd een lage correlatie tussen de groepen beoordelaars gevonden. Deze correlatie nam toe naarmate de beoordelingsmethode meer was gestructureerd. Tevens bleken de correlaties tussen de economie studenten en de inhoudsdeskundige beoordelaars niet drastisch af te wijken van de correlaties tussen de geneeskunde studenten en de huisartsen indien een minimale structurering aanwezig was.

Uit de resultaten van de G-studie bleek dat door elke beoordelaarsgroep met het vrije oordeel, in vergelijking met de gestructureerde beoordelingsmethoden, het meeste onderscheid tussen personen werd aangebracht en de verschillen in de rangordening van de studenten door de beoordelaars groter was. Daarnaast werd voor de algemene error variantie de grootste component gevonden. Voorts bleek dat met de gestructureerde beoordelingsmethoden grotere verschillen in de moeilijkheidsgraad van de casus werden geconstateerd. Bij de gestructureerde beoordelingsmethoden was de inconsistentie van personen over casus (inhoudsspecificiteit) eveneens beduidend groter dan bij het vrije oordeel.
Opvallend was dat bij alle drie de groepen beoordelaars de algemene error variantie bij het vrije oordeel toenam ten koste van de pi-component (de inconsistentie van personen over casus), terwijl bij de gestructureerde beoordelingsmethoden deze afnam ten gunste van de pi-component.

Met betrekking tot de interbeoordelaarsbetrouwbaarheid werd een onverwachte trend gevonden. Bij de huisartsen nam de interbeoordelaarsbetrouwbaarheid toe naarmate de beoordelingsmethode meer was gestructureerd. Bij de geneeskunde en economie studenten was de overeenstemming het grootst bij respectievelijk de ingedikte en uitgebreide criterialijst. Deze resultaten waren niet conform de verwachting dat de overeenstemming tussen de inhoudsdeskundige beoordelaars het grootst zou zijn met de korte antwoordsleutel en de ingedikte criterialijst. De verwachting dat de interbeoordelaarsbetrouwbaarheid bij de economie studenten zou toenemen naarmate de beoordelingsmethode meer was gestructureerd, werd niet bevestigd.

De totale betrouwbaarheid is in een viertal D-studies bestudeerd. Zoals op basis van de grote ware variantie en de lage pi-component bij het vrije oordeel kon worden voorspeld, werd bij deze beoordelingsmethode de hoogste totale betrouwbaarheid gevonden ongeacht de expertise van de beoordelaars en het gehanteerde design. Bij het geneste design waren de verschillen tussen het vrije oordeel en de gestructureerde methoden groter dan bij het gekruiste design. Dit werd veroorzaakt door de lagere interbeoordelaarsbetrouwbaarheid

bij het vrije oordeel.
Tussen de drie gestructureerde beoordelingsmethoden werden slechts minimale verschillen gevonden in de hoogte van de totale betrouwbaarheid, ongeacht de expertise van de beoordelaars. Deze resultaten vormden dus geen bevestiging van de hypothese dat bij de economie studenten de totale betrouwbaarheid zou toenemen naarmate de beoordelingsmethode meer was gestructureerd. Op basis van de resultaten werd eveneens de hypothese verworpen dat zowel de geneeskunde studenten als de huisartsen met de korte antwoordsleutel en de ingedikte criterialijst tot de hoogste totale betrouwbaarheid zouden komen.

Naast de invloed van structurering van beoordelingsmethoden op de interbeoordelaars- en de totale betrouwbaarheid waren de consequenties van de gehanteerde beoordelingsmethode op de betrouwbaarheid van zak/slaag-beslissingen onderwerp van studie. Als gevolg van de verschillen in de gemiddelde score per beoordelingsmethode varieerde de betrouwbaarheid van de genomen zak/slaag-beslissing bij de verschillende cesuren sterk per beoordelingsmethode. Geconcludeerd werd dat bij de bepaling van de zak/slaag-grens rekening moet worden gehouden met de mate van structurering van de beoordelingsmethode.

Uit de vergelijking tussen de gewogen en de ongewogen scores bleek dat bij de ongewogen scores de eerste- en vierdejaars studenten gemiddeld beduidend hogere waarderingen behaalden dan bij de gewogen scores. Bij de zesdejaars studenten was het omgekeerde waarneembaar. Deze resultaten bevestigden de hypothese dat het aanbrengen van positieve weging de validiteit van de beoordelingsmethoden vergroot.
In het correlationeel onderzoek werden lage correlaties tussen de gewogen en de ongewogen scores voor elke beoordelingsmethode gevonden. Daarnaast bleek dat de correlaties tussen de gestructureerde beoordelingsmethoden bij de ongewogen scores lager waren dan bij de gewogen scores. Op basis van deze resultaten werd geconcludeerd dat het aanbrengen van op inhoudelijke gronden aangebrachte positieve wegingsfactoren de inhoudsvaliditeit van de beoordelingsmethoden vergroot.

Wat betreft de waardering van de beoordelaars voor de verschillende beoordelingsmethoden kwam een zeer gedifferentieerd beeld naar voren. Het merendeel van de huisartsen vond het vrije oordeel en de korte antwoordsleutel het meest adequaat. De geneeskunde studenten waren daarentegen van mening dat met de ingedikte en de uitgebreide criterialijst de meest adequate beoordelingen konden worden gegeven. De meeste economie studenten waren van mening dat met de uitgebreide criterialijst de antwoorden van de studenten het meest adequaat beoordeeld konden worden.

Op grond van de onderzoeksresultaten kon geen uniform advies worden gegeven omtrent de meest adequate mate van structurering. Uit de resultaten bleek echter wel dat het vrije oordeel tot betere resultaten leidde dan voorafgaande aan het onderzoek werd verwacht, en dat door een minimale structurering de verschillen tussen leken en inhoudsdeskundige beoordelaars werden teruggedrongen, evenals de verschillen tussen inhoudsdeskundigen onderling. Tevens toonden de resultaten dat leken met een voldoende inhoudelijke instructie in staat zijn een moeilijk aspect als medisch probleemoplossen te beoordelen.
Concluderend kan worden gesteld dat bij de keuze van een beoordelingsmethode met verschillende aspecten rekening dient te worden gehouden, zoals de deskundigheid van de beoordelaars, de verdeling van de antwoorden over de beoordelaars en de lengte van de

toets.

In het onderzoek is een aantal factoren onder controle gehouden, die in de dagelijkse praktijk van het onderwijs niet kunnen worden beheerst. Welke invloed deze experimentele aanpak kan hebben voor de generaliseerbaarheid van de gegevens naar de onderwijspraktijk, wordt in het laatste hoofdstuk beschreven. Daarnaast wordt in dit hoofdstuk een aantal suggesties voor verder onderzoek gedaan.

SUMMARY

Over the past three decades a great number of written clinical simulations have been developed for the purpose of testing problem-solving skills of medical students. Most studies on those instruments, however, had to report poor psychometric characteristics. Especially, the variability in examinee performance from case to case appears to cause problems. More recently developed instruments include a large number of simulations to overcome this problem of content specificity. Frequently these tests use some kind of modified essay question. Such open-answer type questions introduce a hazard to the reliability of the ratings. In many studies low agreement between raters is found. Even among expert-raters variation in ratings appears to be high. A number of studies have suggested that the lack of agreement between raters is due to the quality of the answer-key. The results showed that too much structure as well as too little structure in the answer-key could induce disagreement. Two important questions, however, remain unanswered. In the first place: to what extent should the answer-key be structured? and, in the second place, what is the relationship between the expertise of the rater and the extent of structure in an answer key?
This thesis discusses an experiment in which the influence of the structuring of the answer-key and the expertise of raters on reliability was studied. The impact of positive weighting on reliability and validity was also investigated. At the same time, the validity of the cases and the practical usefulness of several answer-keys was examined.

In accordance with suggestions by Norman et al. (1987) a test consisting of eight short cases related to patient management problems was developed. The cases consisted of patient problems encountered in an average family physician practice and included both physical and psychological complaints. Each case was followed by one to three restricted response questions (a total of 11) which focussed on the quintessence of the problem. The cases were divided into four groups according to the stages of the S.O.A.P. scheme, resulting in two separate cases per stage.
The cases were administered to 98 medical students from three classes (18 first year students, 40 fourth year students and 40 sixth year students). For each question four different scoring methods were developed varying in the extent of structure. With increasing structure the scoring methods were: (1) global judgement method, (2) short answer key, (3) global checklist and (4) elaborated checklist. Positive weights were assigned to the categories in the three structured scoring methods to investigate the influence of weights on resulting scores.
The answers were scored by three groups of raters varying in medical expertise. These groups consisted of 16 general practitioners, 16 fourth year medical students and 16 fourth year economical students (laymen with respect to the medical domain). A questionnaire was constructed to investigate the appraisal of the scoring methods.

The study was conducted in two phases. First the construction of the test materials and the feasibility of the design of the study were tried out in a pilot-study. This resulted in a few alterations and the adjustment of the relative weight factors. Next the main study was carried out.

In a first analysis the discriminant validity of the cases was studied by comparing the mean scores of the three classes of medical students. The mean scores on the test appeared to increase significantly as the students had followed more years medical training. Also, the mean scores decreased as the scoring method was more structured. This tendency was found for the four scoring methods and the three groups of raters. The increase in mean scores as the students had more medical education was interpreted as an affirmation of validity.

To investigate the validity of the scoring methods correlations between the four scoring methods and between the three groups of raters were carried out. and discriminant function analyses were carried out. Discriminant function analyses was also carried out to investigate validity of the scoring methods. The number of correct classifications was taken as measure of validity; the more students correctly classified the more valid the scoring method.

The correlations between the global judgement method and the three structured scoring methods were lower than the correlations between the three structured scoring methods, for all three groups of raters. For the layman raters the correlations between the short answer key and the other structured scoring methods were also lower than the correlations between the global checklist and the elaborated checklist. The correlation between the global checklist and the elaborated checklist varied from .75 (general practitioners) to .87 (medical students).

Correlations between the three groups of raters were estimated per scoring method. Low correlations were found for the global judgement method. Correlations increased as the scoring method was more structured. The correlations between the laymen raters and the general practitioners did not differ from the correlations between the medical students and the general practitioners, if the scoring method was minimally structured. These findings suggest that laymen raters are not inferior to experts.

The discriminant function analysis did reveal no clear differences between the four scoring methods. The percentage of correct classifications ranged from 76% to 80% for the three groups of raters. Therefore it was concluded that from this analysis the validity of all four methods was acceptable, and no method was superior.

Generalizability analyses were carried out in order to investigate the relative contribution of the facets students (p), cases (i) and raters (j), and the interactions between these facets on the total variance. Within each group of raters a totally crossed p x i x j-design was available for each scoring method.

The estimated variance component associated with the person facet as well as the p x j interaction term was the highest for the global judgement method. The overall error component was also larger for the global judgement method compared to the structured scoring methods. For the structured scoring methods larger differences between cases were found. The inconsistency of persons over cases (content specificity problem) was considerably higher for the structured scoring methods compared to the global judgement method. A tendency was found that the overall error component for the global judgement method increased at the cost of the p x i-component for the global judgement method. On

the contrary, the overall error component decreased in favor of the p x i-component for the structured scoring methods.

For the economical students it was expected that the inter rater reliability should increase as the scoring method was more structured. For the medical students as well as the general practitioners the highest agreement between was expected to be found for the short answer key and the global checklist. The results showed another picture. For the economical student highest inter rater reliability was found for the elaborated checklist. There was no systematic tendency that the agreement increased as a result of structuring the scoring methods. The agreement between medical students was highest for the global checklist. In contrast with the expectations, the agreement between general practitioners increased as the scoring method was more structured.

The results of the G-study were used in several Decision-studies (D-study) to estimate a number of reliability-coefficients. In one D-study the inter rater reliability was estimated. In the other D-studies the impact of the number of cases and the number of raters on reproducibility of scores was investigated. Finally, a D-study was carried out to study the reliability of pass/fail decisions, using cut-off scores of 30%, 40%, 50% and 60%. The D-studies were conducted within a norm referenced as well as a criterion referenced approach, with a totally crossed and a nested design.

Highest overall reproducibility was found for the global judgement method. This was primarily caused by the high p-component and the low p x i-component. Between the structured scoring methods only small differences were found. These findings were more extreme in a nested design due to the low inter rater reliability for the global judgement method. These results were not in agreement with the hypothesis that the highest overall reproducibility should be found for the short answer key and the global checklist for the medical students as well as the general practitioners. The hypothesis that for the layman raters the reproducibility should increase as the scoring method was more structured, was also not confirmed.

The influence of structuring on the reliability of pass/fail-decisions was also investigated with a D-study. The reliability of pass/fail-decisions substantially varied for the different passing scores and the four scoring methods, due to the differences in mean scores. It was concluded that passing scores should be related to the scoring method used.

To investigate the impact of weighting on the ranking of students, correlations between weighted and unweighted scores were calculated. Also, the correlations between the scoring methods based on weighted composite scores were compared with the correlations based on unweighted composite scores.

This comparison showed that the mean scores for the first and the fourth year students were higher for the unweighted than for the weighted scores. In contrast, the mean score for the sixth year students was lower for the unweighted scored compared to the weighted scores. These results confirmed the hypothesis that positive weighting increases the (discriminant) validity of scoring methods.

Correlation between weighted and unweighted composite scores was low. Besides,

correlations between the structured scoring methods were higher for the weighted than for the unweighted scores. It was therefore concluded that positive weighting increases the construct validity of scoring methods.

Finally the appreciation for the different scoring methods was investigated by means of a questionnaire. Appreciation for the scoring methods varied with the expertise of the raters. The majority of general practitioners preferred the global judgement method and the short answer key most. The medical students preferred the global checklist and the elaborated checklist. The elaborated checklist was also preferred above the other scoring methods by the layman raters.

On the basis of this study it is not possible to recommend a single best scoring method. The global judgment method was more appropriate than expected. It was found that even a minimally structured scoring method can minimalize the differences between raters. Likewise, the results showed that laymen are capable to produce reliable and valid ratings of medical problem solving, provided they are equipped with appropriate information.

This leads to the conclusion that several aspects have to be taken into account when making a choice for a scoring method in a specific situation. Among them the test length, the availability of expert raters, time needed for the construction of answer keys and the distribution of the answers to the raters.

In an experimental study several factors can be controlled, which is not possible in a normal educational setting. As a consequence, the generalizability of the results for the educational practice should be studied. In the last chapter of the thesis a number of suggestions for further study are proposed.